Nucleic Acids Research, 2004, Vol. 32, Database issue ***D45–D49***
DOI: 10.1093/nar/gkh084

ACLAME: A CLAssification of Mobile genetic Elements

Raphaël Leplae*, Aline Hebrant, Shoshana J. Wodak and Ariane Toussaint

Service de Conformation de Macromolécules Biologiques et de Bioinformatique, Université Libre de Bruxelles, Campus Plaine—CP 263, Boulevard du Triomphe, 1050 Bruxelles, Belgium

Received August 15, 2003; Revised and Accepted October 3, 2003

ABSTRACT

The ACLAME database (http://aclame.ulb.ac.be) is a collection and classification of prokaryotic mobile genetic elements (MGEs) from various sources, comprising all known phage genomes, plasmids and transposons. In addition to providing information on the full genomes and genetic entities, it aims to build a comprehensive classification of the functional modules of MGEs at the protein, gene and higher levels. This first version contains a comprehensive classification of 5069 proteins from 119 DNA bacteriophages into over 400 functional families. This classification was produced automatically using TRIBE-MCL, a graph-theory-based Markov clustering algorithm that uses sequence measures as input, and then manually curated. Manual curation was aided by consulting annotations available in public databases retrieved through additional sequence similarity searches using Psi-Blast and Hidden Markov Models. The database is publicly accessible and open to expert volunteers willing to participate in its curation. Its web interface allows browsing as well as querying the classification. The main objectives are to collect and organize in a rational way the complexity inherent to MGEs, to extend and improve the inadequate annotation currently associated with MGEs and to screen known genomes for the validation and discovery of new MGEs.

INTRODUCTION

The sequencing of the complete genomes of several strains of the same bacterial species brought a new dimension to our understanding of horizontal gene transfer (HGT) in prokaryotes, questioning the very meaning of strains and species [see for instance (1–4)]. The two enterobacteria *Escherichia coli* and *Salmonella* share ~70% of their genes (5). A similar level of divergence is seen between the genomes of two *E.coli* strains, the laboratory strain K12 and the pathogen O157-H7, which differ by as much as 20–30% of their genomes. Strikingly, most of the differences are accounted for by prophages (6,7). Similarly, all of the major gaps in the alignment between the genomes of *Listeria monocytogenes* and *Listeria innocua* correspond to the prophages integrated in the latter (8). The pathogenic *Bacillus anthracis* and its close relative *Bacillus thuringiensis*, the source of Bt toxins, offers another example: their chromosomes are extremely similar and they mostly differ by the nature of the plasmids they host (9).

Phages and plasmids are members of the prokaryotic Mobile Genetic Elements (MGEs), which are central players in mobilizing and reorganizing genes, be it within a given genome (intracellular mobility) or between bacterial cells (intercellular mobility). MGEs are defined DNA sequences of widely varying length (1 to several hundred kb), which often carry the functions that drive their transfer and recombination with the host genome. They are now seen as key players in the reshuffling of genetic material, which in combination with mutations and selection, drive evolution.

Traditionally, MGEs have been classified as bacteriophages, plasmids or transposons. This classification becomes exceedingly obsolete as many chimerical elements are identified as the many types of so called 'genomic islands', which share genes or entire groups of genes. The modular nature of MGEs and the potential for reshuffling between modules has been long recognized, see for instance (10) for the tailed phages, (11) for plasmids. A systematic analysis of MGEs in terms of their modules might therefore be warranted. It is made difficult by the fact that these elements are not adequately annotated in the existing databases, out of which they cannot be easily retrieved. MGE are often referred under the name of the host in which they were identified. As a result, the same (or a closely related) element residing in widely different hosts cannot be readily identified as being related. Complete genome annotations are not particularly helpful either. In the present gene ontologies, the vast majority of MGE gene functions are classified under the three very general 'plasmid', 'phage' or 'transposon' related functional categories.

Using as the basis the idea that MGEs would best be described in terms of a hierarchy of functional modules, at both the protein and DNA levels, we undertook the development of the ACLAME database (http://aclame.ulb.ac.be). The first version presented here, contains information on proteins encoded by genuine prokaryotic 'mobile elements' (phages, plasmids, transposons and other genomic islands), as well as proteins from other sources that are significantly similar to those. It provides a comprehensive classification of these

*To whom correspondence should be addressed. Tel: +32 2 650 5499; Fax: +32 2 650 5425; Email: raphael@scmbb.ulb.ac.be

proteins into families with similar sequences and related functions. The main aims of the ACLAME database are to (i) provide a common framework for the representation of MGEs and their components, (ii) offer tools for facilitating their analysis; (iii) provide a community-wide discussion platform with the goal of deriving a consistent ontology for MGE functions that could then be used to provide a consistent functional annotation of MGEs in all genomes.

DATABASE CONTENT

ACLAME (version α) contains information on 119 DNA bacteriophage genomes, represented by 5069 proteins. Phage sequences were chosen because of their abundance in prokaryotic genomes, their wide size distribution, their very compact genome organization and the resulting occasional overlap of coding sequences and the wide variety of functions they express. All these features were expected to raise most of the technical challenges to be solved in building up the database. All the information used has been downloaded from the NCBI genomes section (http://www.ncbi.nlm.nih.gov/Genomes/index.html), using a completely automatic procedure, which handles tasks such as protein extraction, MGE information retrieval and database cross-referencing. As a first step towards building a comprehensive MGE functional classification, all the proteins have been clustered into families using TRIBE-MCL, a graph-theory-based automatic Markov clustering algorithm that uses sequence measures as input (12). We obtained 437 clusters containing at least three members. These clusters covered a total of 2501 proteins, representing 50% of all analyzed proteins. The remaining proteins were left as singletons or in pairs. In an attempt to build up larger clusters from those, and in order to help in the functional annotation of all clusters, it was deemed useful to search different sequence databases for related proteins and thereby exploit any available functional annotations on those. To that end several searches were performed. In one, individual sequences in each cluster were used to search in Swiss-Prot (13) using 3 iteration Psi-Blast (14) and an E-value threshold of 0.001. The sequences identified and their corresponding functional annotations are stored in dedicated tables linked to the original cluster, which can be queried by the user, or expert annotator. In another, the protein sequences in each cluster were multiply aligned using ClustalW (15) with standard setting. Using these multiple alignments, Hidden Markov Models (HMMs) were built with the HMMER package (http://hmmer.wustl.edu/) (16) and used to screen the NRDB-NCBI (17), Swiss-Prot (13) and SCOP (18) databases for additional sequences more remotely related to those in each cluster. These database search results are stored in dedicated tables also linked to the original clusters and accessible for query and annotation purposes.

FUNCTIONAL ANNOTATION: TOOLS AND DISCUSSION PLATFORM

The first 300 clusters were analyzed manually via the web interface. A functional annotation could easily be assigned to 233 of them based on experimental evidence, similarities in previous annotations and matches in screened protein databases. Clearly, the present functional annotation should be considered as a first draft to be reviewed by the specialized scientific community, using the discussion platform provided on the ACLAME website. A typical example is provided here by cluster 11, which features 24 proteins, some of which are annotated (helicase or putative helicase), but others are not. On the basis of our searches using an HMM in SCOP, NRDB and Swiss-Prot, plus the Psi-Blast searches in Swiss-Prot with individual sequences, the function of the cluster can be unambiguously defined as DNA replication with all its proteins clearly related to helicases. In contrast, cluster 7 comprises 28 proteins, all annotated as having 'unknown' function. Interestingly, some MGEs are represented in this cluster by more than one of their proteins. Knowing that functional redundancy in viruses is rare, it could be interesting to understand why, in the case of cluster 7, some of them have multiple copies of proteins with probably the same function.

Our annotations rely on the use of a slightly extended MultiFun classification scheme (19), which is presently the reference for bacterial genome annotation, and the Gene Ontology (20), whenever a satisfactory definition could be found (see our website for the proposed extensions). A group of experts has been invited to review our annotations and help in improving and extending it with the ultimate aim of developing a consistent and complete annotation scheme applicable to the many different types of MGE. A major concern is dealing with the many inconsistencies currently encountered in the different ontologies. We noticed for instance that the GO definitions for transposase, integrase and site-specific recombination are not correct for prokaryotic organisms. Moreover, this ontology in its present form does not properly consider MGEs that have not been found in eukaryotes.

ACCESS AND INTERFACE

At present, the ACLAME database can be accessed only via a web interface; however, the data and the classification will be stored in a relational database under the MySQL database management system (21), which will be directly and publicly accessible to the scientific community. The web interface allows browsing the MGE genomes currently loaded in the database, their associated proteins and the hosts in which they are found (Fig. 1). The protein classification obtained through the automatic clustering procedure is also accessible (Fig. 2) and the annotation tools are available to registered users.

A Blast search interface has been implemented in order to query the database. Results of sequence or genome analysis can be accessed and retrieved through various file formats (XML, tab delimited, ...).

FUTURE DIRECTIONS

The ultimate goal in developing ACLAME is to be able to define functional modules, which are well-characterized features, found in MGEs, independently of their 'generic identification'. Defining such modules should not only allow the reconstruction of known MGEs but more importantly, should enable their high complexity and the difficulty in identifying them across genomes to be dealt with. For instance, many comparisons between phages originating from related bacteria have been published throughout the

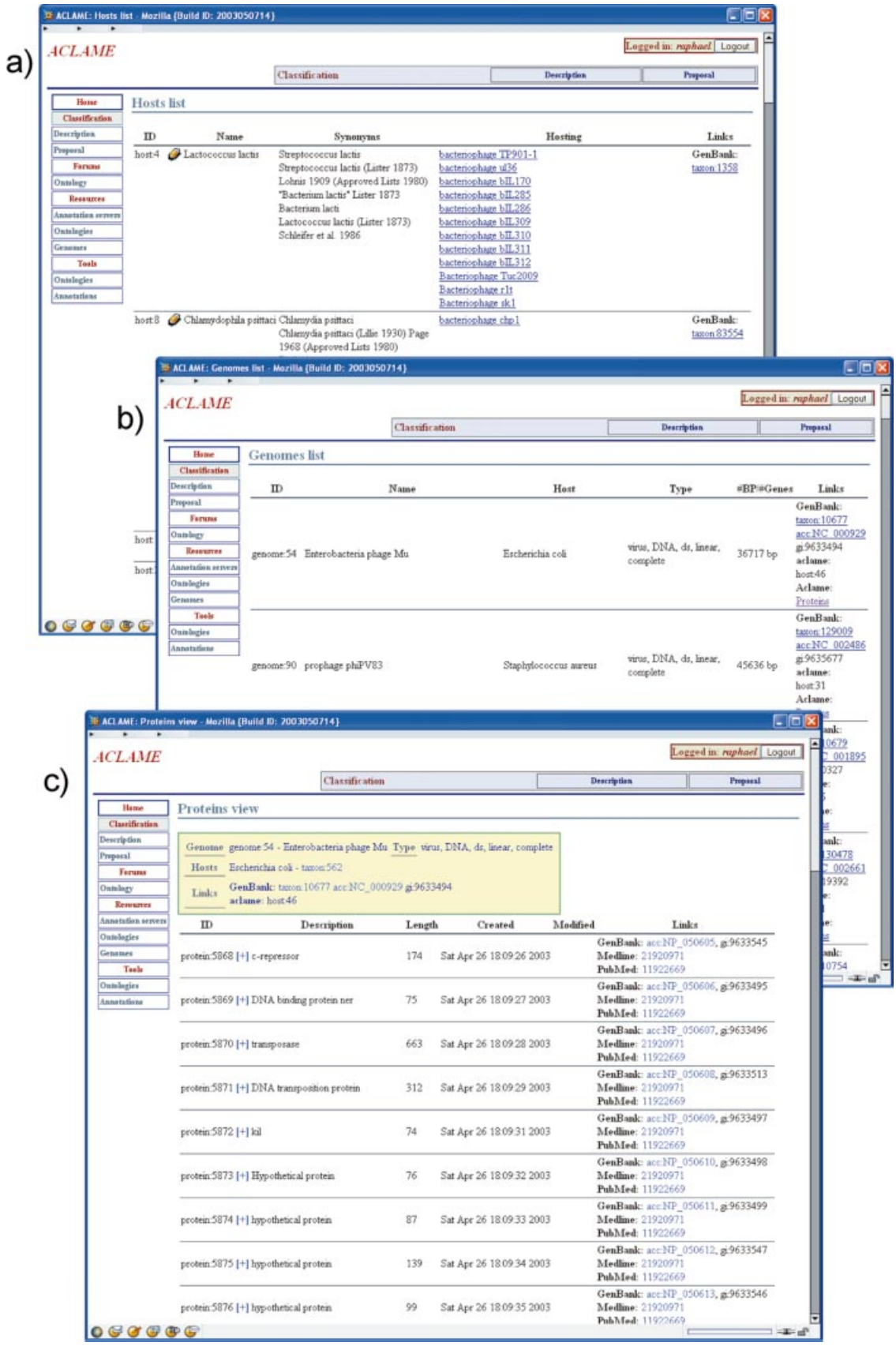

Figure 1. Three views of the MGE genomes loaded in ACLAME. (**a**) The list of hosts where the MGE genomes are found. (**b**) The list of MGE genomes with their major features. (**c**) The list of proteins found in a MGE genome. Links to external databases are provided in all pages.

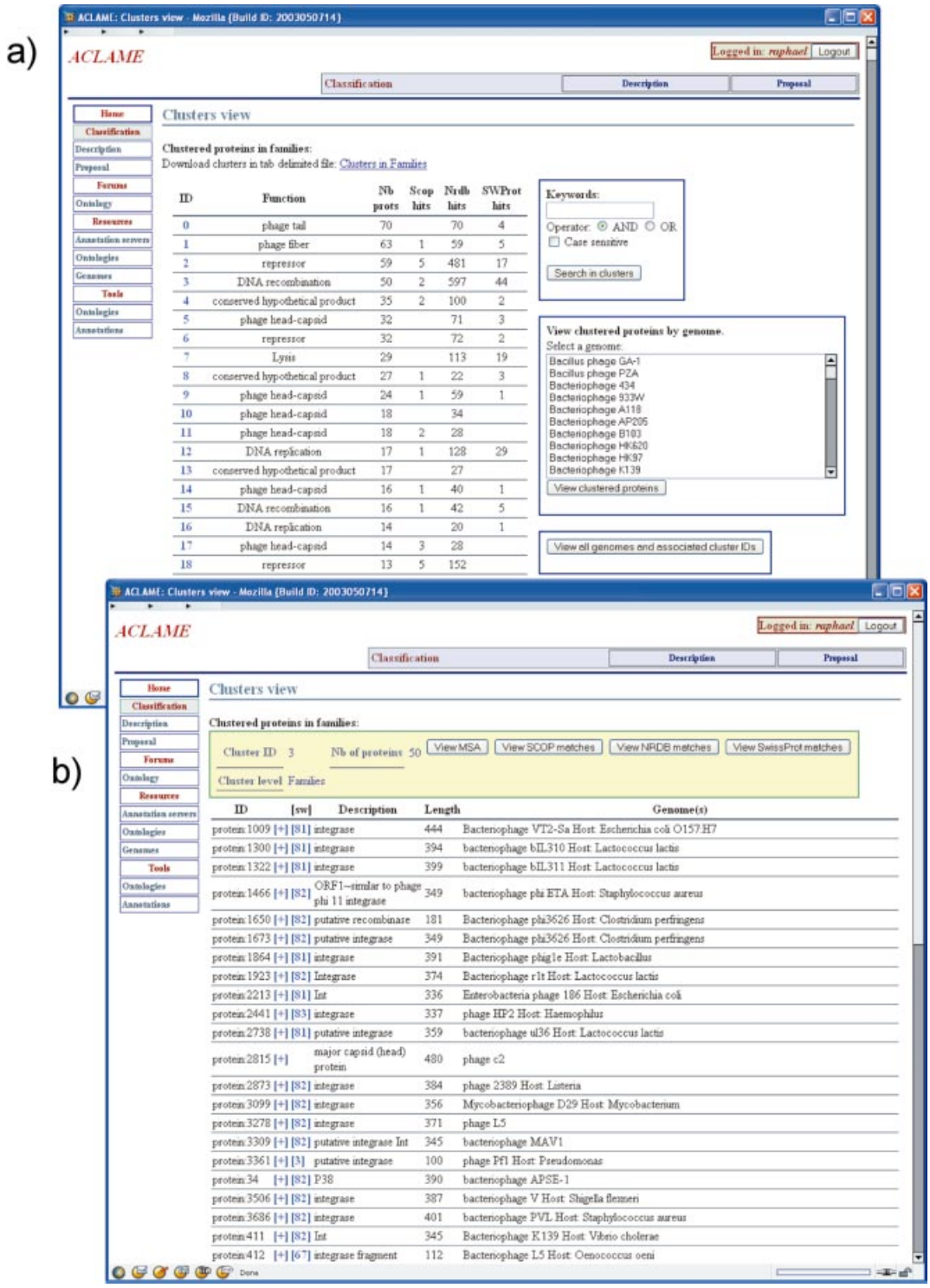

Figure 2. Access to the clustered MGE proteins. (**a**) The list of clusters with the number of proteins found in each of them, the functional annotation and the number of hits found using HMMs in SCOP, NCBI-NRDB and Swiss-Prot. Some tools are provided to further analyze these clusters. (**b**) Content of a cluster with the list of proteins, their original annotation and the MGE genome where they are found. Access to the database searches with Psi-Blast 3 iterations in Swiss-Prot and/or NCBI-NRDB is also provided.

years [see (22) and (23) for the most recent ones]. To our knowledge, only one study so far has covered all available completely sequenced phage genomes (24), but this classification was related to the taxonomy of the phages analyzed, one aspect that we deliberately omitted. Indeed our aim is to traverse the largest spectrum of MGEs and associated protein/DNA sequences, independently of their taxonomy. This we hope will help in cases like for instance the DD-E transposase proteins, to provide a general classification covering the whole range of IS sequences, transposons, conjugative transposons,

transposable phages and pathogenicity islands, all of which encode enzymes of that family. Such a description should better reflect the functional roles and evolutionary history of the considered modules, thereby hopefully allowing the derivation of a taxonomy and ontology that rest on a more rational basis. To achieve this goal, we will continuously update ACLAME with information from newly sequenced MGE genomes and with help of expert knowledge in the scientific community.

ACKNOWLEDGEMENTS

We are grateful to Mik Chandler and Jacques Mahillon for providing the IS sequences and for help in analyzing the clustering results. We are also grateful to Max Mergeay for the support and fruitful discussions related to this project. This work was supported in part by ESA, the European Space Agency, contract ESTEC 16370/02/NL/CK and the Fonds de la Recherche Scientifique Médicale (FRSM). A.T. is Research Director from the Fonds National de la Recherche Scientifique (FNRS).

REFERENCES

1. de la Cruz,F. and Davies,J. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.*, **8**, 128–133.
2. Hacker,J. and Carniel,E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.*, **2**, 376–381.
3. Gogarten,J.P., Doolittle,W.F. and Lawrence,J.G. (2002) Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, **19**, 2226–2238.
4. Lawrence,J.G. (2002) Gene transfer in bacteria: speciation without species? *Theor. Popul. Biol.*, **61**, 449–460.
5. Parkhill,J., Dougan,G., James,K.D., Thomson,N.R., Pickard,D., Wain,J., Churcher,C., Mungall,K.L., Bentley,S.D., Holden,M.T. *et al.* (2001) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, **413**, 848–852.
6. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
7. Perna,N.T., Plunkett,G.,3rd, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
8. Glaser,P., Frangeul,L., Buchrieser,C., Rusniok,C., Amend,A., Baquero,F., Berche,P., Bloecker,H., Brandt,P., Chakraborty,T. *et al.* (2001) Comparative genomics of *Listeria* species. *Science*, **294**, 849–852.
9. Read,T.D., Peterson,S.N., Tourasse,N., Baillie,L.W., Paulsen,I.T., Nelson,K.E., Tettelin,H., Fouts,D.E., Eisen,J.A., Gill,S.R. *et al.* (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature*, **423**, 81–86.
10. Casjens,S., Hatfull,G. and Hendrix,R. (1992) Evolution of the dsDNA tailed-bacteriophage genomes. *Semin. Virol.*, **3**, 383–397.
11. Couturier,M., Bex,F., Bergquist,P.L. and Maas,W.K. (1988) Identification and classification of bacterial plasmids. *Microbiol. Rev.*, **52**, 375–395.
12. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
13. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
15. Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.*, **8**, 189–191.
16. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
17. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
18. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
19. Serres,M.H. and Riley,M. (2000) MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
20. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
21. DuBois,P. (2003) *MySQL.* Sams Developer's Library, pp. 1219.
22. Pedulla,M.L., Ford,M.E., Houtz,J.M., Karthikeyan,T., Wadsworth,C., Lewis,J.A., Jacobs-Sera,D., Falbo,J., Gross,J., Pannunzio,N.R. *et al.* (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell*, **113**, 171–182.
23. Canchaya,C., Proux,C., Fournous,G., Bruttin,A. and Brussow,H. (2003) Prophage genomics. *Microbiol. Mol. Biol. Rev.*, **67**, 238–276.
24. Rohwer,F. and Edwards,R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.*, **184**, 4529–4535.